Software Specification for the NRGL

Introduction

The main functions of digital libraries for storage of any documents, thus also for the repository of grey literature, are retrieving, processing and storage of data, making them accessible and data protection. Typical activities performed in the area of the grey literature lie mainly in the organizational area of document retrievals and also deal with copyright. They do not significantly influence the software selection for the repository of grey literature, in general, the selection criteria are identical with the criteria of software selection for any digital library service. The decisive criteria for the selection of the best bidder will be based on how the SW supports the main functions of the digital library. Additional criteria are aimed to ascertain other useful features, such as user and administrator comfort, and other extra functions, e.g. hardware requirements, operating software platform, availability of services, possibility to deploy the software in more phases and so on. The selection of software for the repository of grey literature or any other digital library must be dealt with proper care because of the digital nature of the library. Therefore the decision about the software selection is the most important.

According to this view the requirements for software for STL NRGL it can be divided into five categories:

- **Content Management of the Digital Library** – This criterion applies to the main functions of the digital library. It includes storage, versioning and content checks (e.g. upon migration of files) of the digital library, linear browsing (lists) as well as hierarchical browsing, metadata administration, various mechanisms to retrieve data (e.g. harvesting – OAI-PHM, web services), interoperability for cooperation with other digital libraries (e.g. Z39.50, OAI-PHM, SRU), support for great volumes of metadata, multilingual search above the metadata and files (full text), Boolean operators and proximity search, support of various formats of files stored – text formats (e.g. DOC, RTF, PDF), images (e.g. TIFF, JPEG, GIF, PNG), presentations (e.g. MS PPT, Adobe PDF and PostScript), structured formats (e.g. HTML and XML), audio and video (e.g. WAV, MP3, AVI, MPEG4, Real audio and video) and support of coding (e.g. ASCII, UNICODE, UTF-8).

- **User interface** – Access via www supporting the main internet browsers, the possibility to customise the user interface for various user roles from the inside as well as outside of the library, multilingual user interface, multilingual presentation of search results and web 2.0 functions.

- **User management** – Classification and user profiles on the basis of their roles (e.g. RBAC), authentication of users on the basis of users' names and passwords, management of IP addresses access, access via proxy, tracking and reporting to evaluate users' behaviour to be used later to improve services of the digital library, authorisation of users (e.g. Shibolet, Eduroam, OpenID) and user activity logs for billing purposes.

- **System Administration** – This criterion applies to the administration of the operating environment of the digital library which can be arduous and demanding. It covers the tools for administration of the metadata, setting of automated mechanisms of data collection, indexing, automated generating of keywords, formatting of reports and so on.

- **Other Requirements** – User and administrator documentation, helpdesk (from the producer / implementer of the system), extra functions for users such as discussion groups, automated alerts of various events (occurrence of new documents meeting specific criteria), RSS exports, user folders and comments on the documents, requirements for hardware and software and so on.

The main process of the NRGL

Content Acquisition – content providers will make the data accessible in various ways, via the web services, the OAI-PHM standard, e-mail, by ftp-transfer, on a data carrier, or submitting them directly. In case that the metadata are retrieved via OAI-PHM contain URL(s) to file(s), also individual files can be downloaded in case it is permitted in the agreement with the provider.

Content Processing – The data provided must be processed, the metadata must be converted to the NRGL format by means of conversion templates, files must be converted to the presentation format (PDF / A), and to prepare the batch for the import to the database with the possibility to check it and upload the metadata to the database, either the presentation form of the data, or the original form of files in various formats as well, and to generate indexes and keywords and to format the search procedures.

Making the Content Accessible – Users access the database via the www interface anonymously as public user, if they access from particular institutional addresses, the system signs them in as anonymous institutional users, or they sign in with user names and passwords which can be obtained in the library system, or their authorisation is done via the systems such as Shibolet, Eduroam or OpenID. Users can browse the contents according to various criteria in a linear way in lists as well as hierarchically in the tree structure, or by the form of Web 2.0 to search above the metadata with the file contents by means of Boolean operators, full text, or using proximity search. They can also store the search results to permanent user boxes (shopping baskets), write comments on the contents in discussion groups and request alerts about events in the collections (occurrence of new records and so on) by e-mail or to their user desktop. The system enables user activity logs for billing purposes, or to make the paid services accessible by means of an interface to another system.

Storage and Protection of the Content – quality checks in respect of the content manager, data duplicity, quality of the metadata, possibility to create queries above the database, checks of primary files for sustainability of the content and so on.

Administration of the System - administration tools to control, configure and administer the whole system including control of operations and their scheduling over the central database and also allows manual intervention (e.g. to shift a started time-consuming task to a later time), administration of user accounts, roles and access rights, automated alerts about various events, possibility to create administrator queries over the database, and back-up system of the data and so on.

General functional specification

1.1. system and databases for storage of bulk number of library records on the basis of defined metadata from various internal as well as external resources and for their fast and efficient search and providing them to the end users via the web interface, supporting the interoperability with other digital repositories
1.2. automated testing of quality and metadata integrity (e.g. length of fields, numeric/ alphanumeric content, obligatory / optional / dependent fields) and providing alerts for administrators and cataloguing staff
1.3. automated recognition of keywords from the text documents on the basis of occurrence of words from a defined dictionary
1.4. conversion of the metadata from proprietary formats to standard formats as well as to the NRGL format, and it allows full configurability of the formats of inputs as well as outputs (system of the structure to describe the input and output formats, similarly as the XSLT for XML). The system shall also have set of templates for conversions from/to the most common and known formats and a possibility to edit these templates via web interface
1.5. formatting of bibliographic metadata for various purposes, e.g. for outputs when performing the search and presentation of documents, to separate the administration of the data contents from the administration of the output design of the templates, possibility to edit them via the web interface
1.6. automated mechanism for data collection (harvesting), supporting common standards, based on OAI-PHM (Open Archive Initiative Protocol for Metadata Harvesting) + OAI-ORE? or OAI-x? and supporting two-way interoperability with other digital libraries, web harvesting (data downloading via the web), web submission (providing the data via the web interface from authorised subjects including conversion of full text documents from various other text formats as image formats), e-mail upload (it allows upload of the data sent by e-mail), support of search in other digital libraries
1.7. indexing system of the metadata, links and full text files and mapping of these indexes to metadata tags to enable faster search in the database, administration of index definitions
1.8. classification system of records in the database according to various criteria for later search (e.g. frequency of word occurrence, tag value, number of views of a record, number of requests for a document and so on)
1.9. check of input data against the existing database with configurable criteria, e.g. to prevent duplicate data storage
1.10. administration tools to manage, configure and administer the whole system including the control of operations and their scheduling over the central database also with the possibility of manual intervention (e.g. shifting of a commenced time-consuming operation to a later time). control of access to the system on the basis of roles (RBAC – Role Based Access Control)
1.11. bulk upload of previously formatted data (XML) to the central database including consistency checks
1.12. user web interface for search, presentations and providing documents with support of various extra functions, such as personal user boxes to store the documents (similar to shopping baskets), setting of various alerts to inform about occurrence of documents with defined keywords, user discussion groups, communication user tools (boards, user comments, rating of the documents and so on)
1.13. search tool with a web interface which allows search according to words including phrases as well as more complex queries with Boolean criteria, a structured presentation of search results (classification according to the types of documents), and to offer alternatives in case it had not been found according to the criteria in search.

Based on the aforesaid requirements, we stipulate these criteria for the purpose of evaluation:

- **Content Management** – tools and procedures supporting the submitting of the contents in the digital library and management of this process of submitting and versioning
- **Content Acquisition** – import and export of the content, support of various document formats
- **Metadata** – support of various metadata formats is important for indexing, submitting and making the contents accessible.
- **Search Support** – it applies to a number of search and browsing functions, search in the metadata, full text search, hierarchical browsing and so on.
- **User Management and Privacy Protection** – user management and privacy protection include administration of passwords, user accounts with access rights as well as retrieval of forgotten passwords.
- **Support of Reports and Search Queries** – this criterion deals with a possibility to evaluate the usage of the digital library and to monitor users' behaviour patterns in order to improve services, and usage of user activity logs for billing purposes
- **Sustainability, Data Protection** – protection of the metadata, consistency and integrity of the database, back-up, support of migration of the metadata
- **Interoperability –** to provide two-way integration with other distributed systems on the metadata level, search as well as retrieval and providing of documents, support of the OAI-PHM, Z39.50
- **User Interface** – this category deals with support of other languages and the possibility to modify the user interface according to various users' needs or different implementations
- **Standards Compliance** – standards are important for sharing and long-term storage of the digital content. It is a wide range of standards from the metadata to interoperability as well as formats of the saved documents
- **Automation Tools** – this category deals with automation tools for acquisition of content, harvesting, generating of the metadata, maintenance and so on.
- **Support, Services** – an important of all software systems. Numerous digital libraries come from the area of Open Source, where this criterion must be taken into account in particular. Important services are: documentation, helpdesk, collection of requirements to improve it, discussion forums and so on.
- **Hardware and software** – hardware requirements for the repository, operation and back-up of the system, securing its availability, supported operating and database systems and so on.

2. Individual criteria

2.1. Content Management

2.1.1. Support of the system for a multiple collections within one installation

2.1.2. Tools for the administrator for parametrization of content submitting (to set the process of import, input formats, process of conversion)

2.1.3. Is it possible to set the page templates for each collection separately?

2.1.4. Definition of access rights to submit the contents on the basis of roles

2.1.5. Configurable access rights to submit the content within various collections (roles for the collections)

2.1.6. E-mail alerts to users about events in the collections and which information can be set

2.1.7. E-mail alerts to administrators about events in the collections and which information can be set

2.1.8. User (individual data supplier) checks of the saved data

2.1.9. It enables the administrator to perform checks of the data provided

2.2. Content Acquisition

2.2.1. Upload of compressed data

2.2.2. Upload of the data from a known address – URL

2.2.3. Bulk upload of data

2.2.4. Bulk upload of the metadata to existing collections

2.2.5. Bulk export / transferability of the content to another system

2.2.6. Possibility to restrict allowed content formats by the administrator

2.2.7. Entity used for upload can contain more files and/or types of files

2.2.8. Text files

2.2.8.1. ASCII

2.2.8.2. Unicode

2.2.8.3. RTF

2.2.8.4. Other

2.2.9. Images

2.2.9.1. JPEG

2.2.9.2. TIFF

2.2.9.3. GIF

2.2.9.4. Other

2.2.10. Presentations

2.2.10.1. MS PPT

2.2.10.2. Adobe PDF

2.2.10.3. Adobe PostScript

2.2.10.4. Other

2.2.11. Structured formats

2.2.11.1. HTML

2.2.11.2. XML

2.2.11.3. SGML

2.2.11.4. Other

2.2.12. Multimedia

2.2.12.1. Wave

2.2.12.2. MP3

2.2.12.3. Real Audio

2.2.12.4. AVI

2.2.12.5. MPEG

2.2.12.6. Real Video

2.2.12.7. Other

2.2.13. Possibility to acquire previous versions of files / records

2.2.14. Possibility to identify changes

2.2.15. Possibility to compare changes

2.3. Metadata

2.3.1. Possibility to upload, make changes and of indexing of accepted contents in real time

2.3.2. Supported formats / metadata standards

2.3.2.1. MARC 21

2.3.2.2. Dublin Core

2.3.2.3. EAD

2.3.2.4. LOM

2.3.2.5. METS

2.3.2.6. MODS

2.3.2.7. VRA CORE Categories

2.3.2.8. Other

2.3.3. Possibility to add / delete items of metadata

2.3.4. Possibility to set default values of the metadata items

2.3.5. Support of Unicode for the metadata

2.4. Search support

2.4.1. Full text search

2.4.1.1. Boolean operators

2.4.1.2. wildcards

2.4.1.3. phrase search

2.4.1.4. proximity search

2.4.1.5. approximate search of similar phrases

2.4.2. Search of descriptive metadata

2.4.2.1. Boolean operators

2.4.2.2. wildcards

2.4.3. Search of selected items of metadata

2.4.4. Browsing of the records according to

2.4.4.1. author

2.4.4.2. title

2.4.4.3. publication date

2.4.4.4. subject

2.4.4.5. collections

2.4.4.6. added / modified fields

2.4.4.7. more items at a time

2.4.5. Categorising of the search results according to

2.4.5.1. author

2.4.5.2. title

2.4.5.3. publication date

2.4.5.4. relevance

2.4.5.5. other criteria

2.5. User Management and Privacy Protection

2.5.1. User passwords are assigned by the system

2.5.2. Users choose the passwords themselves

2.5.3. Functions to retrieve the forgotten passwords

2.5.4. Creation of user accounts

2.5.5. Editing of user accounts

2.5.6. Deletion of users

2.5.7. Restriction of access to the level of fields / objects

2.5.8. Restriction of access to the level of collections

2.5.9. Groups / user roles can be defined

2.5.10. Restriction of access according to the roles

2.5.11. Collections can be adapted to individual roles

2.5.12. Filtering of source IP addresses

2.5.13. Proxy filtering

2.5.14. Access based on credits

2.5.15. Support of encrypting when entering sensitive data

2.5.16. Support of digital signatures

2.6. Support of Reports and Search Queries

2.6.1. The system generates statistics about usage

2.6.2. If yes, which?

2.6.3. Is scheduling possible for reports?

2.6.4. Can reports be edited?

2.6.5. Are there editable templates for reports?

2.6.6. User activity logs for billing purposes

2.7. Security, data protection

2.7.1. Is permanent identification of the documents secured?

2.7.2. Are the identifiers assigned by the system?

2.7.3. Is there a CNRI Handles support?

2.7.4. Does the system support quality checks?

2.7.5. If yes, how?

2.7.6. Is there a proposed procedure how to protect the digital data?

2.7.7. If yes, describe in brief.

2.8. Interoperability

2.8.1. Support of OAI-PHM – metadata harvesting

2.8.2. Support of Z39.50 protocol

2.8.3. E-mail upload

2.8.4. Web upload

2.8.5. Web harvesting (OAI-ORE, OA-X …)

2.8.6. Search protocol Dienst

2.8.7. Search protocol SDLIP

2.9. User interface

2.9.1. Is the user interface customizable?

2.9.2. Is it possible to attach a customised header / footer to the static / dynamic pages?

2.9.3. Support of multilingual user interface within one installation of the system

2.10. Support of standards – summary

2.10.1. Structured data - HTML, XML, SGML …

2.10.2. Metadata - Dublin Core …

2.10.3. Text - ASCII, Unicode, RTF …

2.10.4. Images - JPG, TIFF, GIF …

2.10.5. Presentations – MS PowerPoint, Adobe PDF, Adobe PostScript …

2.10.6. Multimedia - wav, mp3, avi, mpeg, real audio, real video …

2.11. Automation Tools

2.11.1. system for metadata input

2.11.2. Generating of indexes

2.11.3. Generating of HTML pages

2.11.4. Generating of reports

2.12. Support, Services

2.12.1. Documentation, manuals

2.12.2. Discussion forums, e-mail groups

2.12.3. Error reporting

2.12.4. Collection of requirements for development

2.12.5. Support – helpdesk

2.13. Hardware, operating software

2.13.1. Hardware requirements for the server

2.13.2. Hardware requirements for client stations

2.13.3. Supported server operating systems

2.13.4. Supported client operating systems

2.13.5. Supported databases (MySQL, Oracle …)